Musical Instrument Identification using Multiscale Mel-frequency Cepstral Coefficients

Sturm, Bob L.; Morvidone, Marcela; Daudet, Laurent

Published in:
Proceedings of the European Signal Processing Conference

Publication date:
2010

Document Version
Accepted author manuscript, peer reviewed version

MUSICAL INSTRUMENT IDENTIFICATION USING MULTISCALE MEL-FREQUENCY CEPSTRAL COEFFICIENTS

Bob L. Sturm, Marcela Morvidone, and Laurent Daudet†*

Department of Architecture,
Design and Media Technology
Aalborg University Copenhagen
Lautrupvang 15, 2750
Ballerup, Denmark
email: bst@imi.aau.dk

*Universidad Tecnologica Nacional
Facultad Regional Buenos Aires
Departamento de Ingenieria Electrica
Campus: Mozart 2300 (C1407IVT)
C.A.B.A. Buenos Aires Argentina,
email: morvidone@gmail.com

†Institut Langevin (LOA),
UMR 7587
Université Paris Diderot
10, rue Vauquelin, 75231 Paris
Cedex 05, France
email: laurent.daudet@espci.fr

ABSTRACT

We investigate the benefits of evaluating Mel-frequency cepstral coefficients (MFCCs) over several time scales in the context of automatic musical instrument identification for signals that are monophonic but derived from real musical settings. We define several sets of features derived from MFCCs computed using multiple time resolutions, and compare their performance against other features that are computed using a single time resolution, such as MFCCs, and derivatives of MFCCs. We find that in each task — pairwise discrimination, and one vs. all classification — the features involving multiscale decompositions perform significantly better than features computed using a single time-resolution.

1. INTRODUCTION

The cepstrum, and Mel-frequency cepstral coefficients (MFCCs), provide very successful features in tasks of speaker verification [1, 2], and speech recognition [3], by the fact that the human voice can be modeled extremely well over short time scales by filtering a wide-band periodic source (glottal impulses) by a small-order linear time-invariant all-pole system (resonances in the throat, mouth, and nose). In practice, MFCCs are computed using a single time resolution, typically 30 ms spaced every 10 ms for speech. However, the human voice is a relatively well-behaved signal compared with other manners of sound production, for instance, plucked strings and percussion in music. Nonetheless, MFCCs and similar features have shown moderate success for musical signals in tasks such as fingerprinting, e.g., [4], and instrument identification, e.g., [5], although their wider application to polyphonic musical signals appears limited [6, 7]. A general problem in using MFCCs for tasks of identification, however, is that signals can contain mixtures and a variety of phenomena that occur over many time-scales. Computing the cepstrum of musical signals using a single time-resolution is probably suboptimal in the sense that it cannot distinguish between these different phenomena. One approach to incorporating time-domain information into discriminating features is the use of derivatives of the MFCCs [5]. This still uses a single time-domain resolution, however, even though one can integrate the features over time-scales longer than the analysis window.

In [8] we propose a novel approach to incorporating time-domain information into MFCC-like features by first decomposing a signal by a greedy iterative descent method of sparse approximation using a multiresolution time-frequency dictionary of Gabor atoms [9]; then finding the distribution of energy in the signal as a function of atom scale and modulation frequency; and then reducing redundancy of the feature space by approximately decorrelating its dimensions using a discrete Cosine transform (DCT). We applied these features to simple tasks of instrument discrimination and classification in a database [10] consisting of monophonic recordings of real instruments that are extracted from real performance contexts, i.e., the violin and cello samples have double and triple stops; there are multiple notes in the piano and guitar samples; there are extended techniques in the trumpet; and the recordings are in real reverberant spaces. For this database and these specific tasks, we found that the features produced by a much more simple approach — DCTs of combined MFCCs evaluated over multiple scales — are more effective for these tasks even though the promise of sparse approximation over a multiresolution dictionary is source separation with respect to the stationarity of phenomena. In other words, we predicted that the more complex approach with sparse approximation could bridge the problems associated with computing MFCCs for signals having a variety of time-scale phenomena. In this article, we investigate more thoroughly the benefits of the simpler approach, and compare their performance against other proposed features that attempt to combine time-domain information, e.g., delta MFCCs, in the context of musical instrument identification.

2. MEL-FREQUENCY CEPSTRAL COEFFICIENTS

We now review the calculation of MFCCs over short time scales before discussing how we incorporate information over multiple scales. Given a real discrete sequence $x[n]$ defined for $0 \leq n \leq N-1$, and its discrete Fourier transform (DFT), $\hat{x}[k] = DFT\{x[n]\}$, and assuming $|\hat{x}[k]|$ to be nonzero everywhere, the *real cepstrum* of $x[n]$ is [3]

$$c_x[l] \stackrel{\Delta}{=} DFT^{-1}\{\log|\hat{x}[k]|\} = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \log|\hat{x}[k]|\, e^{j2\pi kl/N} \quad (1)$$

for $l = 0, 1, \ldots N/2+1$. Considering that $x[n]$ is an audio signal, we would like to have a compact and perceptually-based description of its spectral characteristics. We can do so by replacing the magnitude DFT in (1) by the energy observed in frequency bands that are exponentially-spaced according to perceived pitch. Such a relationship between frequency and pitch is given by *Mel-frequency scaling*, which maps Mel fre-

quency $\phi \geq 0$ to Hz f: $f(\phi) = 700(e^{\phi/1127} - 1)$. Thus, we may construct a filterbank of L filters with center frequencies linearly spaced in Mels, and substitute the energy of its outputs into (1) to find a perceptually-relevant spectral description of $x[n]$.

Many variations exist for the filterbanks used [2], but here we use $L = 48$ overlapping bands with triangular magnitude responses weighted such that each has equal area, and together the span a bandwidth of $[0, 9614]$ Hz. (This represents an increased bandwidth compared with our work in [8], which used 40 overlapping bands spanning a bandwidth of $[133, 6854]$ Hz.) Each filter here ($l = 1, 2, \ldots, 48$) is given by

$$\hat{h}_l[k] \triangleq \begin{cases} 0, & 0 \leq kF_s/N < f_c(l-1) \\ a_l \frac{kF_s/N - f_c(l-1)}{f_c(l) - f_c(l-1)}, & f_c(l-1) \leq kF_s/N < f_c(l) \\ a_l \frac{kF_s/N - f_c(l+1)}{f_c(l) - f_c(l+1)}, & f_c(l) \leq kF_s/N < f_c(l+1) \\ 0, & f_c(l+1) \leq kF_s/N \leq Fs \end{cases} \quad (2)$$

where F_s is the Nyquist sampling rate, $f_c(0) = 0$, $f_c(49) = 9614$ Hz, the band-dependent magnitude factors are given by

$$a_l \triangleq \begin{cases} 0.015, & 1 \leq l \leq 14 \\ \frac{2}{f_c(l+1) - f_c(l-1)}, & 15 \leq l \leq 48. \end{cases} \quad (3)$$

and the center frequency of the lth band is given by

$$f_c(l) \triangleq \begin{cases} 66.\overline{66}l, & l = 1, 2, \ldots, 14 \\ 1073.4(1.0711703)^{(l-14)}, & l = 15, 16, \ldots, 48. \end{cases} \quad (4)$$

The MFCCs of $x[n]$ are defined as the discrete cosine transform of the energies of the L filterbank outputs, i.e.,

$$cc_x[m] \triangleq \beta_L(m) \sum_{l=1}^{L} \log \left(\sum_{k=0}^{N-1} \left| \hat{x}[k] \hat{h}_l[k] \right| \right) \cos \left[\frac{m\pi}{L} \left(l - \frac{1}{2} \right) \right] \quad (5)$$

for $0 \leq m < L$, where the normalization factor is defined

$$\beta_L(m) \triangleq \begin{cases} \sqrt{1/L}, & m = 0 \\ \sqrt{2/L}, & m > 0. \end{cases} \quad (6)$$

Typically in speech processing [3], only the first $M = 13$ coefficients are kept excepting the term at $m = 0$ since it is related only to the signal energy. For music signals, it is common to use more coefficients, e.g., $M = 20$ [4, 7].

For non-stationary signals MFCCs are evaluated over short time-scales using overlapping sliding windows. Time-localized, or short time, MFCCs are given by

$$cc_x[m, p] \triangleq \beta_L(m) \sum_{l=1}^{L} \left(\sum_{k=0}^{P-1} \log \left| \hat{x}[k, p] \hat{h}_l[k] \right| \right) \times \cos \left[\frac{m\pi}{L} \left(l - \frac{1}{2} \right) \right] \quad (7)$$

for $0 \leq m < L$, and where the length-P DFT of $x[n]$ localized over the time region $[p, p+s)$ is defined

$$\hat{x}[k, p] \triangleq \frac{1}{\sqrt{P}} \sum_{n=0}^{P-1} x[n+p] w[n] e^{-j2\pi kn/P}, 0 \leq k \leq P-1. \quad (8)$$

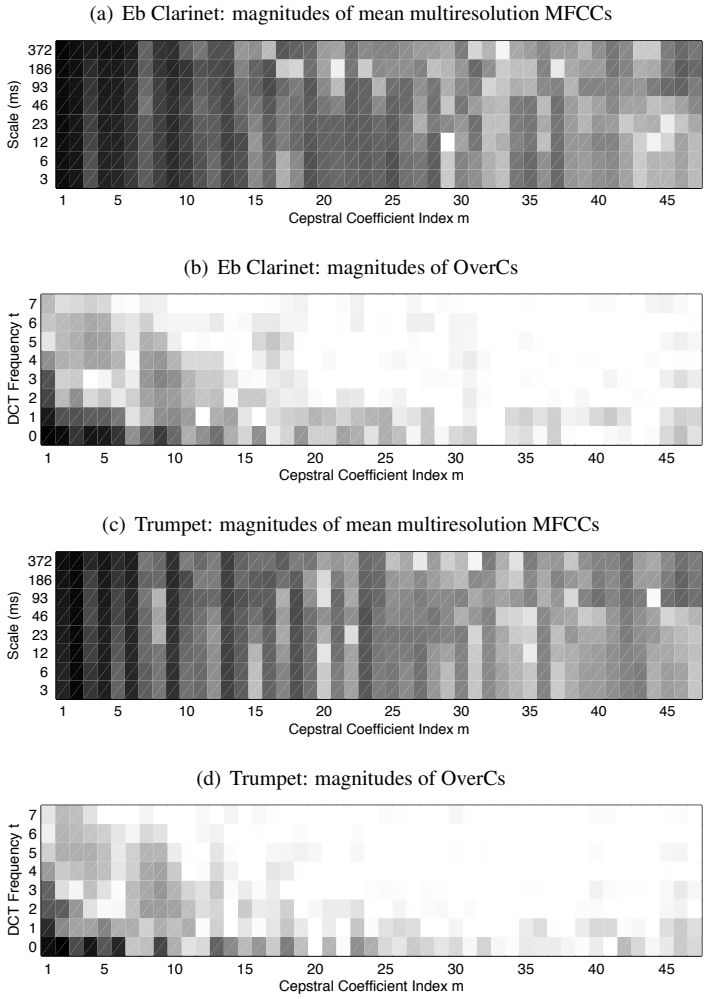

Figure 1: Magnitudes of mean multiresolution MFCCs, $\overline{cc_x}[m, S(s)]$ in (12), and of OverCs, $\zeta_x[m, t]$ in (13), for two instruments playing a chromatic scale from C5 to B5.

for time shifts $0 \leq p < N - s$, and a real window $w[n]$ with support $s \leq P$. In speech and music processing, typical window lengths are 10 – 100 ms, with hop sizes of half their duration. Zeropadding can be applied, i.e., $P > s$, to interpolate the frequency domain samples. Finally, we compute an instantaneous derivative of MFCCs features, or ΔMFCCs, by

$$\Delta_x[m, p] = cc_x[m, p] - cc_x[m, p-1] \quad (9)$$

and its second derivative, or ΔΔMFCCs, by

$$\Delta\Delta_x[m, p] = \Delta_x[m, p] - \Delta_x[m, p-1]. \quad (10)$$

These features describe how the MFCCs change between subsequent frames of data, and can be likened to frame-rate spectral flux.

3. INCORPORATING SCALE INFORMATION

While the definition of time-localized MFCCs in (7) involves a single time-scale over which the signal is observed, we proposed in [8] to compute multiscale MFCCs-like features over several time-scales using a method of sparse approximation with a multiscale Gabor dictionary, or compiling

$S(s)$ index	s (samples/ms)	Δ_p (samples/ms)	Δ_f (Hz)
1	128/2.9	64/1.5	43.1
2	256/5.8	128/2.9	43.1
3	512/11.6	256/5.8	43.1
4	1024/23.2	512/11.6	43.1
5	2048/46.4	1024/23.2	21.5
6	4096/92.9	2048/46.4	10.8
7	8192/185.8	4096/92.9	5.4
8	16384/371.5	8192/185.8	2.7

Table 1: Multi-scale MFCC parameters for signals with a sampling rate of $F_s = 44.1$ kHz: scale (window size) s, time resolution (window hop) Δ_p, and DFT frequency resolution Δ_f (zero-padding added to interpolate frequency-domain samples — which is how we have small scale atoms with a frequency resolution 43.1 Hz).

MFCCs computed with different frame sizes. Since in [8] we focus on the former approach, here we look more closely at the latter. We create the set of time-localized MFCCs, $\{cc_x[m,p,S(s)], S(s) \in \mathcal{S}\}$, where $S(s)$ just maps the scale s to an index in the set $\mathcal{S}$, using the time and frequency resolutions in Table 1. Now define the set

$$\mathcal{C}_{S(s),\varepsilon} \stackrel{\Delta}{=} \{cc_x[m,p,S(s)] : cc_x[0,p,S(s)] > \varepsilon \geq 0\} \qquad (11)$$

which is the collection of length-L short-time MFCCs using the window scale s localized at times when the signal has energy greater than $\varepsilon \geq 0$ (to avoid problems when there is no signal). For each scale index $S(s)$, we compute the mean MFCCs over the set $\mathcal{C}_{S(s),\varepsilon}$ to give *mean multiresolution MFCCs*

$$\overline{cc_x}[m,S(s)] \stackrel{\Delta}{=} \frac{1}{|\mathcal{C}_{S(s),\varepsilon}|} \sum_p cc_x[m,p,S(s)]. \qquad (12)$$

Note that each scale index $S(s)$ expresses the short-time MFCCs over $x[n]$ averaged over the entire signal using a scale s; and each cepstral index expresses how a particular mean MFCC changes as a function of the analysis scale used. Figure 1(a, c) show examples of this feature for two signals created by musical instruments playing an ascending and chromatic scale from C5 to B5 over a duration of 32 seconds (Clarinet), and 97 seconds (Trumpet).

Since there is redundancy across scales, we uncouple the values in each MFCCs coefficient by performing a discrete Cosine transform in the scale direction. This creates features we call *OverCs* [8]:

$$\zeta_x[m,t] \stackrel{\Delta}{=} \beta_{|\mathcal{S}|}(t) \sum_{\sigma \in \mathcal{S}} \overline{cc_x}[m,\sigma] \cos\left[\frac{t\pi}{|\mathcal{S}|}\left(\sigma - \frac{1}{2}\right)\right] \qquad (13)$$

defined for $0 \leq t < |\mathcal{S}|$. Figure 1(b, d) show the OverCs for the same musical signals.

4. SIMULATIONS

These new features — mean multiresolution MFCCs and OverCs — are of course in a higher-dimensional space than are mean MFCCs, and so we select four subsets of 20 coefficients each to use in our classification experiments. In addition, we construct three other 20-dimensional features that do not use time-scale information at all, but two of which include delta features. For all features, we set $\varepsilon = 0.1$ so as to avoid frames that have very little signal energy, i.e., we only use features from signal frames that have zeroth cepstral coefficients greater than ε. The detailed set of features we test in tasks of identification are the following:

1. *Mean MFCCs* (MFCCs): from $\overline{cc_x}[m,S(s)]$ in (12), the first 20 coefficients ($m = 1,2,\ldots,20$) for $s = 46.4$ ms with hop of 23.2 ms;
2. *Mean MFCCs with mean* Δ*MFCCs* (MFCCsΔ): from $\overline{cc_x}[m,S(s)]$ in (12), the first 10 coefficients for $s = 46.4$ ms with hop of 23.2 ms; and the mean ΔMFCCs in (9) of the first 10 coefficients at same scale and hop;
3. *Mean MFCCs with mean* Δ*MFCCs and mean* $\Delta\Delta$*MFCCs* (MFCCs$\Delta\Delta$): from $\overline{cc_x}[m,S(s)]$ in (12), the first 8 coefficients for $s = 46.4ms$ with hop of 23.2 ms; and the mean ΔMFCCs in (9) of the first 6 coefficients at same scale and hop; and the mean $\Delta\Delta$MFCCs in (10) of the first 6 coefficients at same scale and hop;
4. *Mean multiscale MFCCs* (MSMFCCs): from $\overline{cc_x}[m,S(s)]$ in (12), the first 10 coefficients for $s = 46.4$ ms with hop of 23.2 ms; and the first 5 coefficients for a scale of $s = 2.9$ ms with hop of 1.5 ms; and the first 5 coefficients for a scale of $s = 371.5$ ms with hop of 185.8 ms;
5. *OverCs* (OverCs1): from $\zeta_x[m,t]$ in (13), using parameters in Table 1, the first 20 coefficients with $t = 0$;
6. *OverCs* (OverCs2): from $\zeta_x[m,t]$ in (13), using parameters in Table 1, the first 14 coefficients with $t = 0$; the first 6 coefficients with $t = 1$;
7. *OverCs* (OverCs3): from $\zeta_x[m,t]$ in (13), using parameters in Table 1, the first 12 coefficients with $t = 0$; the first 5 coefficients with $t = 1$; the first 3 coefficients with $t = 2$;

Our logic in choosing the subsets of MSMFCCs features is that the first 10 coefficients of the middle row of Fig. 1(a,c) describe the mean power spectral shape over an average window size, while the first 5 coefficients from each of the largest and smallest window sizes provides additional information on how the mean power spectrum changes with these extremely different time-scales. When we take the DCT of the MSMFCCs in the scale direction, thus producing the OverCs in Fig. 1(b,d), the coefficients from the 0th DCT frequency represent a feature closest to the mean MFCCs (feature 1) [8]. By combining with these coefficients at higher scale frequencies, we aim to provide information on how the mean cepstral coefficients vary with time-scale.

Our music signal database, which has been used in other work, e.g., [8, 10], consists of 2,755 five-second monophonic signals excerpted from real musical recordings with no overlap between segments. These are recordings of real music played in real environments, some of which are from commercial CDs, and are not isolated single notes. Furthermore, many of the segments include extended performance techniques, and non-traditional styles. Each of the seven instrument classes — Cl: clarinet, Co: cello, Gt: guitar, Ob: oboe, Pn: piano, Tr: trumpet, Vl: violin — contains signals from five different sources, i.e., different performer, instrument, composition, recording, etc.

To identify an unknown instrument we use support vector machines (SVM) with a radial basis function [8, 10, 11]. We find the best parameters using a grid search method [8, 12].

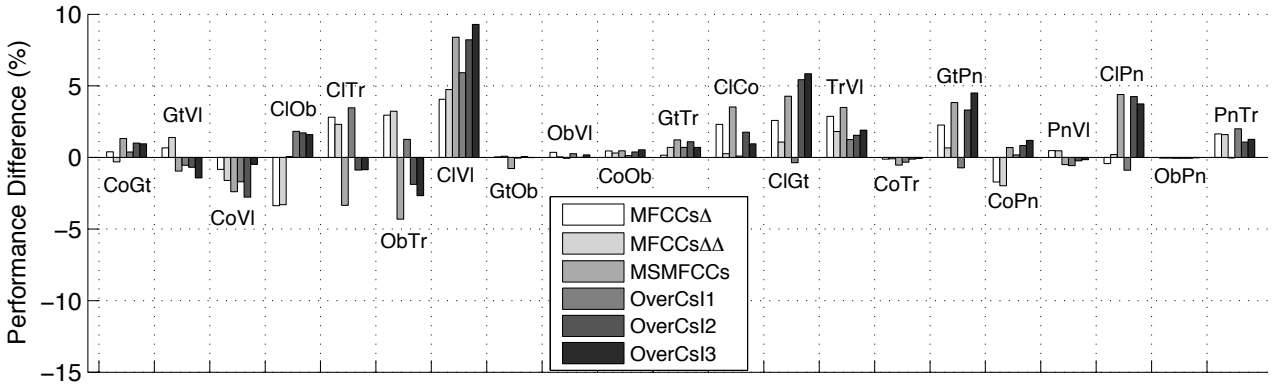

Figure 2: Mean correct instrument discrimination rates for features of all pairs relative to that of the mean MFCCs at a single scale.

To perform instrument identification we train each SVM using five-fold cross validation of data randomly selected from four different sources for each instrument class. We then identify the testing data selected randomly from the remaining sources of each instrument class. We never include features from the same source in both the training (and grid search) and testing data so as to avoid biasing the classifier performance.

Figure 2 shows for each instrument pair the mean gains made in discrimination for the different features with respect to mean MFCCs. We compute these means using independent 100 trials with 49 realizations randomly selected from each instrument class. We see that large gains are made when using scale information (MSMFCCs and OverCs) for ClVl, ClGt, GtPn, and ClPn. The only pairs that suffered from including scale information are GtVl and CoVl. The top portion of Table 2 shows the overall discrimination rates for each of the features. Here we see that OverCs2 and OverCs3 perform the best, with an ANOVA analysis showing that the results from the two features have a p-value of 0.03, i.e., they are likely from different distributions. The likelihood is $p < 10^{-4}$ that the classification results using either OverCs2 or OverCs3 are from the same distributions of the classification results using the other features. We find that the MSMFCCs and MFCCsΔ features do not perform significantly better ($p \approx 0.38$) than the features OverCs2 and OverCs3.

The mean confusion tables of our instrument classification simulations are shown in Table 4. We compute these means using independent 100 trials with 49 realizations randomly selected from each instrument class. The numbers in bold show the highest scores. The OverCs features have five of the highest scores, with MSMFCCs only performing best for classification of Gt, and MFCCsΔ only performing best for Tr classification. Actually, with $p \approx 0.68$, it appears the results from using MFCCsΔ and MFCCsΔΔ come from the same distribution in classifying Tr. For five of the seven instruments (Cl, Gt, Ob, Pn, and Vl), the best performing feature involving scale (MSMFCCs and OverCs) performs significantly better ($p < 0.007$) than the best performing feature that does not involve scale. For Co, the performance of OverCs2 is not significantly better than MFCCs and MFCCsΔΔ ($p < 0.5$). For Tr, the performance of MFCCsΔ and MFCCsΔΔ are significantly better than any of the multi-scale features ($p < 10^{-8}$). For all features, Co is often misclassified as Vl; but Gt is misclassified as Cl less often when using features incorporating scale. Finally, the lower portion of Table 2 shows the mean classification rates over all instruments. Here we see OverCs2 performs significantly better than all other features ($p < 0.014$).

Discrimination	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Mean	94.38	95.21	94.93	95.27	94.96	95.53	**95.66**
Stan. dev.	4.7	4.23	4.53	4.93	4.46	4.48	4.25
Classification	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Mean	79.85	80.78	81.04	84.0	81.88	**84.69**	82.72
Stan. dev.	5.56	5.06	5.38	5.09	6.29	4.61	5.27

Table 2: Correct instrument discrimination and classification rates for each feature: (1) Mean MFCCs. (2) MFCCsΔ. (3) MFCCsΔΔ. (4) MSMFCCs. (5) OverCs1. (6) OverCs2. (7) OverCs3.

5. CONCLUSION

In this paper, we have explored more thoroughly the effectiveness of combining MFCCs features computed over various time scales in the context of musical instrument identification, thus building upon our previous work [8]. We find that in most of the cases we tested the multiscale MFCCs features perform significantly better than features that do not incorporate information from multiple time-scales, e.g., mean MFCCs computed over a single time scale with delta features. This provides further evidence pointing to the effect that we can improve to a large extent the performance of musical instrument classifiers that use MFCC-like features by incorporating features computed over multiple time-scales, and not just by incorporating how features change over time.

Our current work involves using a feature selection strategy for finding the best subset of the multiscale features MSMFCCs and OverCs that provide the best performance in instrument identification tasks, and the effectiveness of these features in classifying instruments in polyphonic signals. We are also looking at the implications of our work for other tasks in music signal processing that use short-term but mono-resolution features.

6. ACKNOWLEDGMENTS

LD acknowledges partial support from Agence Nationale de la Recherche (ANR), project DREAM (ANR-09-CORD-006-04).

REFERENCES

[1] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 4, pp. 430–451, Aug. 2004.

[2] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various mfcc implementations on the speaker verification task," in *Proc. Int. Conf. Speech Computer*, Patras, Greece, Oct. 2005, vol. 1, pp. 191–194.

[3] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Upper Saddle River, New Jersey, 1993.

[4] M. Casey, C. Rhodes, and M. Slaney, "Analysis of minimum distances in high-dimensional musical spaces," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 1015–1028, July 2008.

[5] C. Joder, S. Essid, and G. Richard, "Temporal integration for audio classification with application to musical instrument classification," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 17, no. 1, pp. 174–184, Jan. 2009.

[6] J-.J. Aucouturier, B. Defreville, and F. Pachet, "The bag of frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *J. Acoust. Soc. America*, vol. 122, no. 2, pp. 881–891, Aug. 2007.

[7] J. H. Jensen, M. G. Christensen, D. P. W. Ellis, and S. H. Jensen, "Quantitative analysis of a common audio similarity measure," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 693–703, May 2009.

[8] M. Morvidone, B. L. Sturm, and L. Daudet, "Incorporating scale information with cepstral features: experiments on musical instrument recognition," *Patt. Recgn. Lett.*, 2010 (in press).

[9] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, Academic Press, Elsevier, Amsterdam, The Netherlands, 3rd edition, 2009.

[10] S. Essid, *Classification automatique des signaux audio-fréquence: reconnaissance des instruments de musique*, Ph.D. thesis, Université Pierre et Marie Curie, Paris 6, Paris, France, Dec. 2005.

[11] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," Software available at: http://www.csie.ntu.edu.tw/∼cjlin/libsvm, 2001.

[12] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Tech. Rep. http://www.csie.ntu.edu.tw/ cjlin, National Taiwan University, Taiwan, China, May 2009.

MFCCs	Cl	Co	Gt	Ob	Pn	Tr	Vl
Cl	75.76	0.59	3.12	5.37	1.45	8.00	5.71
Co	2.31	81.12	2.29	0.10	0.00	0.00	14.18
Gt	14.20	1.35	78.35	0.00	5.94	0.00	0.16
Ob	5.20	0.10	0.00	88.43	0.00	6.27	0.00
Pn	6.24	1.10	12.94	0.00	77.84	1.88	0.00
Tr	9.98	0.00	0.00	8.10	1.59	76.84	3.49
Vl	11.73	6.14	0.24	0.02	0.04	1.20	80.61

MFCCsΔ	Cl	Co	Gt	Ob	Pn	Tr	Vl
Cl	71.51	0.84	6.02	9.47	0.92	7.98	3.27
Co	3.29	80.22	2.45	0.16	0.04	0.00	13.84
Gt	12.61	1.12	78.69	0.00	7.55	0.00	0.02
Ob	10.10	0.35	0.00	84.63	0.00	4.92	0.00
Pn	8.80	0.16	10.31	0.00	79.71	1.02	0.00
Tr	5.55	0.00	0.00	4.88	0.61	**86.22**	2.73
Vl	6.35	8.18	0.20	0.04	0.04	0.76	84.43

MFCCsΔΔ	Cl	Co	Gt	Ob	Pn	Tr	Vl
Cl	71.86	0.45	5.57	8.31	1.00	9.31	3.51
Co	2.88	81.39	2.41	0.14	0.06	0.00	13.12
Gt	15.10	1.18	77.51	0.00	6.14	0.00	0.06
Ob	9.82	0.41	0.00	84.90	0.00	4.84	0.04
Pn	8.24	0.35	9.00	0.00	81.10	1.16	0.14
Tr	5.39	0.02	0.00	5.57	0.41	85.94	2.67
Vl	5.94	8.22	0.20	0.20	0.04	0.78	84.61

MSMFCCs	Cl	Co	Gt	Ob	Pn	Tr	Vl
Cl	80.98	1.04	1.24	7.41	0.24	8.16	0.92
Co	2.18	80.47	1.22	0.12	0.27	0.02	15.71
Gt	3.18	1.18	**89.76**	0.00	3.16	0.00	2.71
Ob	7.06	0.18	0.00	88.82	0.00	3.94	0.00
Pn	1.18	0.86	11.94	0.00	85.29	0.73	0.00
Tr	14.88	0.00	0.00	6.45	1.02	74.55	3.10
Vl	2.24	8.67	0.65	0.02	0.04	0.22	88.14

OverCs1	Cl	Co	Gt	Ob	Pn	Tr	Vl
Cl	79.12	0.90	5.00	3.22	2.16	6.18	3.41
Co	2.88	79.80	2.98	0.16	0.06	0.00	14.12
Gt	11.37	1.29	80.18	0.00	5.49	0.00	1.67
Ob	5.39	0.08	0.00	89.53	0.00	5.00	0.00
Pn	9.18	0.39	10.86	0.00	78.92	0.65	0.00
Tr	7.80	0.16	0.00	6.18	1.22	81.76	2.88
Vl	7.73	6.94	0.88	0.22	0.06	0.33	83.84

OverCs2	Cl	Co	Gt	Ob	Pn	Tr	Vl
Cl	**83.78**	1.35	1.90	3.94	0.22	8.20	0.61
Co	2.24	**81.92**	1.67	0.39	0.02	0.00	13.76
Gt	3.41	0.86	88.92	0.00	4.57	0.00	2.24
Ob	3.53	0.31	0.00	**90.12**	0.06	5.94	0.04
Pn	1.61	1.08	9.92	0.00	**86.63**	0.73	0.02
Tr	12.59	0.02	0.00	7.47	1.61	75.92	2.39
Vl	2.86	10.37	0.51	0.00	0.00	0.76	85.51

OverCs3	Cl	Co	Gt	Ob	Pn	Tr	Vl
Cl	83.53	0.65	2.04	4.43	0.24	8.53	0.57
Co	1.82	77.59	1.76	0.14	0.00	0.02	18.67
Gt	6.80	1.82	83.22	0.00	3.65	0.00	4.51
Ob	6.69	0.47	0.04	83.12	0.04	9.59	0.04
Pn	1.90	0.41	11.80	0.02	84.37	1.51	0.00
Tr	11.00	0.22	0.00	7.14	1.06	77.10	3.47
Vl	2.71	6.31	0.39	0.00	0.00	0.49	**90.10**

Table 3: Mean confusion tables from 100 independent trials for one-vs.-all instrument classification. Left-hand column is instrument class presented; top rows are instrument class selected by classifier.